Econometric Analysis

Ch. 6 Additional Single-Equation Topics

Ryuichi Tanaka

6.1 Estimation with Generated Regressors and Instruments

6.1.1 OLS with Generated Regressors

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \gamma q + u$$
$$q = f(\mathbf{w}, \boldsymbol{\delta})$$

q : unobservable, f(.,.) : known function

$\mathbf{w}$: observables, $\boldsymbol{\delta}$: unknown parameters

$\hat{\boldsymbol{\delta}}$: consistent estimator

$\hat{q}_i = f(\mathbf{w}_i, \hat{\boldsymbol{\delta}})$: generated regressor (by Pagan 1984)

OLS is consistent if u is uncorrelated with all xs

Standard Error of Generated Regressors

◆In general, we need to modify **standard errors and test statistics**

If $E[\nabla_\delta f(\mathbf{w}_i, \hat{\boldsymbol{\delta}})' u] = \mathbf{0}$ and $\gamma = 0$, no need to modify (in the asymptotics)

Ex. Assuming $E[u \mid \mathbf{x}, \mathbf{w}] = \mathbf{0}$, $E[\nabla_\delta f(\mathbf{w}_i, \hat{\boldsymbol{\delta}})' u] = \mathbf{0}$ is satisfied in usual case

Test of $H_0 : \gamma = 0$

Under homoskedasticity, usual t-test can be applied

Under heteroskedasticity, use robust standard errors for the test statistics

If $H_0 : \gamma = 0$ is rejected, use modified standard errors and test statistics!

6.1.2 2SLS with Generated Instruments

$y = \mathbf{x}\boldsymbol{\beta} + u, \mathbf{x} = (1, x_2, \mathrm{L}\ , x_K)$

$\mathbf{z}$: 1xL IV, $E(\mathbf{z}'u) = \mathbf{0}$

generated instruments: $\hat{\mathbf{z}}_i = \mathbf{g}(\mathbf{w}_i, \hat{\boldsymbol{\lambda}})$

consistency is usually guaranteed

If $\hat{\boldsymbol{\lambda}}$ is $\sqrt{N}-consistent$ and $E[\nabla_\lambda \mathbf{g}(\mathbf{w}_i, \hat{\boldsymbol{\delta}})' u] = \mathbf{0}$,

usual test stat can be used

Ex. Assuming $E[u \mid \mathbf{w}] = \mathbf{0}$, $E[\nabla_\delta f(\mathbf{w}_i, \hat{\boldsymbol{\delta}})' u] = \mathbf{0}$ is usually satisfied

2SLS with generated instruments is easier to use than OLS with generated regressors

6.1.3 Generated Instruments and Regressors

$y = x\beta + \gamma f(w,\delta) + u, E(u \mid z, w) = 0$

δ: estimates from the first stage

$H_0 : \gamma = 0$ can be tested using t-stat with the estimates from 2SLS of

$y_i = x_i\beta + \gamma\hat{f}_i + error$ with $(z_i, \hat{f}_i)$ as IVs

6.2 Some Specification Tests

6.2.1Test of Endogeneity

$y_1 = z_1\delta_1 + \alpha_1 y_2 + u_1$

y_1 : dependent variable

y_2: potentially endogenous variable

z_1: exogenous variables $(1 \times L_1)$

z : exogenous variables$(1 \times L)$ $(z_1 \subset z, E(z'u_1) = 0)$

Assuming order condition and rank condition, test endogeneity of y_2

Regression-based Hausman Test

Consider LPof y_2 on $\mathbf{z}$: $y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2, E(\mathbf{z}'v_2) = \mathbf{0}$

Since $E(\mathbf{z}'u_1) = \mathbf{0}$, if$y_2$ is endogenous $(E(y_2u_1) \neq 0)$, then $E(v_2u_1) \neq 0$

Hence, it is enought to test $E(v_2u_1) = 0$

Consider LPof u_1 on v_2: $u_1 = \rho_1 v_2 + e_1$

Note $\rho_1 = E(v_2u_1)/E(v_2^2)$, $E(v_2e_1) = 0$, and $E(\mathbf{z}'e_1) = 0$

Hence, "y_2 is exogenous if $\rho_1 = 0$"

To test $\rho_1 = 0$, plugging $u_1 = \rho_1 v_2 + e$ in the population model, we have

$y_1 = \mathbf{z_1}\boldsymbol{\delta_1} + \alpha_1 y_2 + \rho_1 v_2 + e_1$

Since e_1 is uncorrelated to $(\mathbf{z_1}, y_2, v_2)$ under the null hypothesis, t-test from OLS is OK

Regression-based Hausman Test (cont.)

However, v_2 is unobserved (need to obtain)

How to obtain $\hat{v}_2$: estimate the first stage $y_2 = z\pi_2 + v_2$ by OLS

$\hat{v}_2 = y_2 - z\hat{\pi}_2$

Summary

(1) obtain $\hat{v}_2$

(2) estimate $y_1 = z_1\delta_1 + \alpha_1 y_2 + \rho_1\hat{v}_2 + \text{error}$ by OLS consistently

(3) test $H_0: \rho_1 = 0$ with usual t-stat

(4) if $H_0: \rho_1 = 0$ is rejected, then y_2 is endogenous

6.2.2 Testing Overidentifying Restrictions

If more than minimum numbers of Ivs are available, we can check if they are uncorrelated to the error term in order to see the validity of these IVs

$y_1 = \mathbf{z_1}\boldsymbol{\delta_1} + \mathbf{y_2}\boldsymbol{\alpha_1} + u_1$

y_1: dependent variable, $\mathbf{y_2}$: potentially endogenous variables $1 \times G_1$

$\mathbf{z_1}$: exogenous variables $1 \times L_1$,

$\mathbf{z} = (\mathbf{z_1}, \mathbf{z_2})$: exogenous variables $1 \times L$ $(L=L_1+L_2)$

If $L_2 > G_1$, the model is over-identified

(Assume order condition and rank condition)

Regression-based Hausman Test

(1) estimate $y_1=z_1\delta_1+y_2\alpha_1+u_1$ by 2SLS and obtain $\hat{u}_1=y_1-z_1\hat{\delta}_1-y_2\hat{\alpha}_1$

(2) regress $\hat{u}_1$ on all IVs (z) and obtain R_u^2

(3) under the assumptions $E(z'u_1)=0$ and homoskedasticity, $NR_u^2 \sim \chi^2_{Q_1}$

where $Q=L_2-G_1$ is the #of over-identification

Basic Idea: If all IVs are exogenous ($E(z'u_1)=0$), R_u^2 must be small because IVs (z) do not explain $\hat{u}_1$

If reject the null hypothesis ($E(z'u_1)=0$), some of z may be inappropriate as IV need reconsider the choice of IVs!

6.2.3 Testing Functional Form

Check if we need higher order terms and/or interaction term

If all explanatory variables are exogenous, we can test by F-test or LM-test if the coefficients on these variables are different from zero (This method may not be practical when the number of higher order terms is large)

Ramsey's (1969) RESET: $y = \mathbf{x}\boldsymbol{\beta} + u, E(u \mid \mathbf{x}) = 0$

Idea : under the assumption $E(u \mid \mathbf{x}) = 0$,

u is uncorrelated to any function of $\mathbf{x}$

Ramsey's RESET

(1) obtain $\hat{\beta}$ by OLS and construct $\hat{y}_i=x_i\hat{\beta}$ and $\hat{u}_i=y_i-x_i\hat{\beta}$

(2) check if $\hat{u}_i$ is correlated with $\hat{y}_i^2, \hat{y}_i^3, \hat{y}_i^4$

(if the population model is nonlinear, $\hat{u}_i$ is correlated with $\hat{y}_i^2, \hat{y}_i^3, \hat{y}_i^4$)

How to check 1: add $\hat{y}_i^2, \hat{y}_i^3, \hat{y}_i^4$ to $y=x\beta+u$, estimate by OLS, and test if all the coefficients on $\hat{y}_i^2, \hat{y}_i^3, \hat{y}_i^4$ are zero by F-test (F(3,N-K-3)) (under homoskedasticity)

How to check2: regress $\hat{u}_i$ on $x_i, \hat{y}_i^2, \hat{y}_i^3, \hat{y}_i^4$ and test if all the coefficients on $\hat{y}_i^2, \hat{y}_i^3, \hat{y}_i^4$ are zero by LM-test ($NR_u^2 \sim \chi_3^2$)

6.2.4 Testing for Heteroskedasticity

$H_0 : E(u^2 \mid \mathbf{x}) = \sigma^2$ (homoskedasticity)

$H_1 : E(u^2 \mid \mathbf{x})$ depoends on $\mathbf{x}$ （heteroskedasticity）

- 分散均一性を帰無仮説におく H_0: $\mathrm{Var}(u/x_1, x_2, \ldots, x_k) = \sigma^2$

この仮説は次のものと同値 H_0: $E(u^2/x_1, x_2, \ldots, x_k) = E(u^2) = \sigma^2$

- u^2 と x_j の関係が線形$u^2 = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + v$との仮定の下で、上の仮説は次のように書ける

H_0: $\delta_1 = \delta_2 = \ldots = \delta_k = 0$

The Breusch-Pagan 検定

$y = \beta_0 + \beta_1 x_1 + \mathrm{L} \ + \beta_K x_K + u$をOLS推計し、

$\hat{u} = y - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \mathrm{L} \ - \hat{\beta}_K x_K$から$\hat{u}^2$を作る

$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \mathrm{L} \ + \delta_K x_K + v$を推計し、帰無仮設をF検定（またはLM 検定）する

F検定量 : $\dfrac{R^2/k}{(1-R^2)/(n-k-1)}$: $F(k, n-k-1)$

The White 検定

- The Breusch-Pagan 検定は、いかなる線形の分散不均一性も探知できる
- The White 検定は、二乗項や交差項を含めることで非線形の分散不均一性も探知するので、より一般的な検定
- The Breusch-Pagan 検定同様、F検定（やLM検定）を用いて帰無仮設を検定する

White検定の別形式

- OLSから得た予測値 $\hat{y}$ をすべての説明変数 x の関数とみなすと、$\hat{y}^2$ はすべての説明変数 x の二乗項と交差項と考えることができる
- ゆえに残差二乗を $\hat{y}$ と $\hat{y}^2$ に回帰し、その R^2 を使ってF検定量（およびLM検定量）を作ることができる
- 帰無仮設は$\hat{y}$ と $\hat{y}^2$ の係数がゼロという簡単なものになる

Test

Exogenous Explanatory Variables Case

$y = \mathbf{x}\boldsymbol{\beta} + u, E(u \mid \mathbf{x}) = 0$

$H_0 : E(u^2 \mid \mathbf{x}) = \sigma^2$, $H_1 : E(u^2 \mid \mathbf{x})$ depends on $\mathbf{x}$

Idea : Consider $1 \times Q$ vector function $\mathbf{h(x)}$

Under $H_0 : E(u^2 \mid \mathbf{x}) = \sigma^2$, $Cov(u^2, \mathbf{h(x)}) = 0$ for any $\mathbf{h(x)}$

$u_i^2 = \delta_0 + \mathbf{h}_i\boldsymbol{\delta} + v$ (assume rankVar(h_i)=Q)

Under $H_0 : E(u^2 \mid \mathbf{x}) = \sigma^2$, since $E(v_i \mid \mathbf{h}_i) = E(v_i \mid \mathbf{x}_i) = 0$, then $\boldsymbol{\delta} = \mathbf{0}$ and $\delta_0 = \sigma^2$

$\Rightarrow$ F- or LM-test $\mathrm{H}_0 : \boldsymbol{\delta} = \mathbf{0}$

Test (cont.)

However, under the null $H_0 : E(u^2 \mid \mathbf{x}) = \sigma^2$, v_i does not follow the normal distribution because $v_i = u_i^2 - \sigma^2 \geq -\sigma^2$

Moreover, $E(v^2 \mid \mathbf{x})$ must be a constant for usual F- and LM -test

homokurtosis assumption: under H_0, $E(u^4 \mid \mathbf{x}) = \kappa^2$(constant)

Using OLS residuals $\hat{u}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$, estimate $\hat{u}_i^2 = \delta_0 + \mathbf{h}_i\boldsymbol{\delta} + v_i$

With R_c^2(centered R-square), $NR_c^2 \sim \chi_Q^2$ (Q is the order of h_i)

Case with endogenous variables

$\mathbf{z}$: IV such that $E(u \mid \mathbf{z}) = 0$ (assume the rank condition)

$H_0 : E(u^2 \mid \mathbf{z}) = \sigma^2$

$\hat{u}_i^2$: by 2SLS residual, and estimate $\hat{u}_i^2 = \delta_0 + \mathbf{h}(\mathbf{z}_i)\boldsymbol{\delta} + v_i$

Under the assumption $Cov(\mathbf{x}, u \mid \mathbf{z})$, test $\boldsymbol{\delta} = 0$ by LM-test

6.3 Single-Equation Methods under Other Sampling Schemes
6.3.1 Pooled Cross Sections over Time

i.n.i.d. (independent, not identically distributed)
ex. Repeated cross-section data

Panel data: **same individuals are repeatedly recorded**
i.n.i.d.: different individuals are recorded

Solution: add time dummies to the Population model!

An example of usage of i.n.i.d. data: Deference in deferences (D-in-D)

Two periods: 1, 2
Two groups: control group A and treatment group B

Ex. (unexpected) change in property tax in between period 1 and 2
control group: no change in property tax
treatment group: change in tax

Evaluate the effect of tax reform : $y = \beta_0 + \delta_0 d2 + \beta_1 dB + \delta_1 d2 \cdot dB + u$

$d2, dB$: dummy variables

δ_1 : the effect of tax reform

OLS estimator of δ_1(D-in-D estimator): $\hat{\delta}_1 = (\bar{y}_{B,2} - \bar{y}_{B,1}) - (\bar{y}_{A,2} - \bar{y}_{A,1})$

$\hat{\delta}_1$ is a consistent estimator (if the change is exogenous)

Example 6.5

Purpose: estimate the effect of an increase in the cap on weekly earnings on the length of time that an injured worker receives workers' compensation

treatment group: high-income workers who were at the cap
control group: low-income workers

Result : high-income worekrs stayed on compensation longer by 19 % due to the increase in the cap

6. 3. 2 Geographically Stratified Samples

$y_{is} = \mathbf{x_{is}}\boldsymbol{\beta} + \mathbf{z_s}\boldsymbol{\gamma} + q_s + e_{is}$, s : *stratum*

$E(e_{is} \mid \mathbf{X}_s, \mathbf{z}_s, q_s) = 0$ for all i and s

$\mathbf{X}_s$: explanatory variables included in s

If interested in $\boldsymbol{\beta}$, estimate $y_{is} = \alpha_s + \mathbf{x_{is}}\boldsymbol{\beta} + e_{is}$ by OLS

If interested in $\boldsymbol{\gamma}$, estimating $y_{is} = \mathbf{x_{is}}\boldsymbol{\beta} + \mathbf{z_s}\boldsymbol{\gamma} + q_s + e_{is}$ by OLS does not guarantee consistency

6.3.3 Spatial Dependence

When cross section units is large relative to population (e.g., state level data), spatial correlation tends to happen

If correlation occurs only through explanatory variables, there is no problem in estimation

If correlation occurs through unobservables, consistency is guaranteed by OLS, but the asymptotics is complicated

6.3.4 Cluster Samples

Cross-section observation correlation

When random sampling many clusters with small size, use usual OLS and 2SLS

Ex.　peer effect in neighborhoods (or families)
Town is cluster level（towns are randomly sampled）
Each town contains small number of students

Ex. Households are randomly sampled
Each household contains small number of household members (siblings)

Asymptotics: increase the number of clusters holding cluster size fixed

PS 6

- 6.1 (testing overidentification)
- 6.3 (test for endogeneity)
- 6.6 (test for heteroskedasticity)
- 6.7 (specification test)
- 6.9 (d-in-d)